LeGeDe - towards a corpus-based lexical resource of spoken German

Möhrs, Christine ; Meliss, Meike ; Batinić, Dolores

Originally published at:
Möhrs, Christine; Meliss, Meike; Batinić, Dolores (2017). LeGeDe - towards a corpus-based lexical resource of spoken German. In: eLex 2017, Leiden, 19 September 2017 - 21 September 2017. Lexical Computing, 281-298.

LeGeDe – Towards a corpus-based lexical resource of spoken German

Christine Möhrs[1], Meike Meliss[2], Dolores Batinić[3]

[1] Institut für Deutsche Sprache, P.O. Box 101621, D-68016 Mannheim

[2] Institut für Deutsche Sprache, P.O. Box 101621, D-68016 Mannheim

[3] Institut für Deutsche Sprache, P.O. Box 101621, D-68016 Mannheim

E-mail: moehrs@ids-mannheim.de, meliss@ids-mannheim.de, batinic@ids-mannheim.de

Abstract

This paper gives an insight into some basic concepts for a corpus-based lexical resource of spoken German, which is being developed by the project "The Lexicon of Spoken German" (Lexik des gesprochenen Deutsch, LeGeDe) at the "Institute for the German Language" (Institut für Deutsche Sprache, IDS) in Mannheim. The focus of the paper is on first ideas of semi-automatic and automatic resources that assist the quantitative analysis of the corpus data for the creation of dictionary content. The work is based on the "Research and Teaching Corpus of Spoken German" (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK).

Keywords: spoken German, corpus linguistics, internet lexicography, lexicology

1. Introduction

The purpose of the project "Lexicon of Spoken German" (Lexik des gesprochenen Deutsch, LeGeDe), which started in September 2016 at the "Institute for the German Language" (Institut für Deutsche Sprache, IDS) in Mannheim, is to build an electronic lexical resource for spoken standard German based on the empiric data of the "Research and Teaching Corpus of Spoken German" (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK[1]). FOLK is the largest corpus of spoken German in interactions (202h/1,95 Mio. tokens; DGD version 2.8) and is made available via the "Database for Spoken German" (Datenbank für Gesprochenes Deutsch, DGD[2]); cf. Schmidt, 2014a/2014b, 2016.

LeGeDe is a third-party funded project [3] of the Leibniz Association (Leibniz Competition 2016, Funding line 1: Innovative projects[4]). For a period of three years (from 1 September 2016 to 31 August 2019) the project will be working on the creation a lexical resource of spoken German.

---

[1] Information about FOLK: http://agd.ids-mannheim.de/folk.shtml.

[2] URL to the DGD-Website: http://dgd.ids-mannheim.de.

[3] Applicants of the project: Annette Klosa, Arnulf Deppermann, Stefan Engelberg, Thomas Schmidt (IDS Mannheim).

[4] For more information about the competition and the funded projects, please go to: http://www.leibniz-gemeinschaft.de/en/about-us/leibniz-competition/projekte-2016/funding-line-1/.

The project is a cooperation of two departments of the IDS in Mannheim: the Department of Pragmatics and the Department of Lexical Studies. The team consists of researchers with different research backgrounds: lexicographers (especially researchers with special focus on electronic lexicography), corpus linguists, and researchers with a special focus on conversational analysis.

The aim of the project is twofold: (1) to develop a lexicographic resource for spoken German (language area: Germany) by benefiting from the methods of corpus-linguistics, and (2) to find an optimal solution for presenting this type of language resource by exploring and extending the possibilities offered by its digital form. The lexicographic resource of spoken German is to be designed in a dynamic (extendible) manner, and it is intended to integrate multi-modal information, such as corpus-based audio-examples and transcriptions for each entry. Hence, compiling such a resource is challenging both from the lexicographic perspective as well as from the point of view of data modelling. In the long term, the resource will be integrated into the dictionary portal OWID[5], which has been developed at the IDS in Mannheim (Online-Wortschatz-Informationssystem Deutsch; eng.: Online vocabulary system of the German language). It will cover, in an exemplary fashion, lexical units and properties typical for spoken German as it is used in conversations in private and institutional contexts.

Modern lexicographic resources of German are usually (and mainly) based on written language represented in large electronic text corpora (e.g. monolingual German dictionaries such as Duden-online, DWDS or *elexiko*). Characteristics of spoken German, especially with regard to the lexicon, are not described in great detail in these dictionaries (cf. Meliss 2016); see the discussion in section 5 on this aspect. LeGeDe is the first project that aims to identify the peculiarities of language in an interactional context in a systematic way (cf. section 5). We are aware of only one similar project focusing on interjections in spoken Danish (cf. Hansen/Hansen 2012) and another one being currently developed for Slovenian (cf. Verdonik & Sepesy Maučec 2017).

The present paper is subdivided into six sections. The subject area of the project is presented in section 2. In section 3, the project's data basis is described. We will present aspects of the quantitative corpus analysis in section 4 and of the data analysis in section 5. The paper concludes in section 6 with final remarks and comments on the further project objectives.

[5] URL to the OWID-Website: www.owid.de.

2. Phenomena of interest

We concentrate on those phenomena which we can characterize as "standard" – in the sense that we intend not to consider dialects (such as Bavarian), sociolects (such as adolescent language) or idiolects. Our interest is mainly directed to those phenomena of spoken German that are used more frequently, or in a different manner than in written German (for example regarding meaning or function in verbal interaction). A selection of phenomena that are to be dealt with in the project is listed in Table 1.

Phenomena of interest (selection)	
Verbs	*ich dachte* (tempus), *guck* (imperative), *meinste* (complementation patterns), Ich *kann* kein Deutsch (modal verbs in absolute use), *geht* (spec. semantics 3rd person) etc.
Word borrowings	German language varieties: *öko*[logisch], *wo* (as a relative pronoun) etc.; Anglicisms: *okay*, *cool*, *fuck* etc. (frequency, groups of speakers, gramm. integration, phonetic realization etc.)
Word formation	*rum-*, *rein-*, *rauf-*; *mega-, super-, sau-, ober-*; *-mäßig (randalemäßig), -i (Hirni)* etc.
Partial synonyms	*kriegen/bekommen/erhalten, gucken/ schauen/sehen; Auto/Karre/Kutsche* etc.
Conversation words	*eben, jein, hä, tss, pf, ups, hoppla* etc.; *gut, richtig, genau, sicher, einfach* etc.
Patterns	*guck mal, alles klar, einen drauf machen* etc.

Table 1: Some phenomena of interest and selected examples

The table provides a rough guide on phenomena and specific lexical units, which should be assigned to the respective phenomena. These areas are also identified as interesting phenomena in research literature (e.g. Schwitalla, 2012; Deppermann, 2005/2007; Fiehler 2016) and in previous studies on spoken German (Imo, 2007; Günthner, 2016; Deppermann et al. (eds.), 2017). With the help of the analysis of corpus evidence the phenomena are to be examined more closely and the candidates should be defined by means of frequency-oriented and competence-based examinations. This should make it possible to draw a clear picture of the relevant phenomena areas, following both a corpus-based and a competence-oriented methodology.

3. Corpus material

We base our research on FOLK that primarily addresses researchers from the fields of conversation analysis and corpus linguistics and comprises conversations from different interaction domains, such as institutional and private conversations, game interactions, table talk, etc. Since the data is annotated on multiple levels (meta

information about speakers, interactions and word forms; cf. Westpfahl & Schmidt, 2016), FOLK provides a reliable basis for a study of interactional phenomena of spoken language, towards which our analysis is mainly directed. Schmidt (2014a) describes its aims as follows:

"[FOLK] has […] set itself the aim of building a corpus of German conversations which:

a) covers a broad range of interaction types in private, institutional and public settings,
b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches,
c) is transcribed, annotated and made accessible according to current technological standards,
d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage." (Schmidt 2014a: 383)

By today, a set of data comprising approximately 202h of recordings and close to 1.95 million transcribed tokens has been completely processed in the FOLK corpus and has been published via the DGD.

Private interaction	**interactions**	**hours**	**tokens**
e.g. coffee table conversation, telephone conversation, conversation on a holiday trip, student everyday conversation, conversation during breakfast, conversation among friends, etc.	89	84:25	864208
Interaction in school/university / at the workplace (non-private/non-public)			
e.g. oral exams at a university, shift change at a hospital, driving school conversation, meeting in an economic company, classroom observation, conversation during a regular meeting, etc.	117	67:53	604121
Public interaction			
mediation talks, panel discussion	6	25:26	237707
Other interaction domains			
maptasks, biographic interviews, interview, ethnographic interviews	47	24:27	246123

Table 2: Interaction domains and examples (selection) in FOLK
(status as of 17.05.2017; cf. also Schmidt 2014a: 383)

FOLK contains transcripts as well as audio and video material on spoken German in interaction. The composition of the corpus can be observed in Table 2. Figure 1 shows the distribution of all tokens over the entire corpus with respect to major interaction domains.

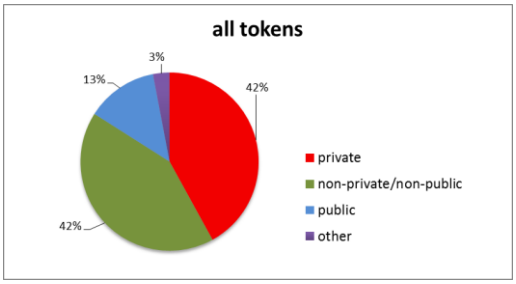

Figure 1: Major interaction domains in FOLK

The list of these different conversations (cf. Table 2) shows the broad diversity of interaction domains covered by FOLK. FOLK's special feature is to document spoken German in spontaneous interaction. This distinguishes it from most other oral corpora in the DGD (see for example the corpus "Deutsche Standardsprache: König-Korpus" which includes reading texts, in particular excerpts from the German Grundgesetz; cf. Schmidt, 2014b: 1451). After the creation of an individual account, the access to the DGD is free of charge for research and teaching purposes. This makes the data base, with which the LeGeDe project works, transparent to the scientific public. Nevertheless, one aspect with regard to FOLK is not to be neglected: Even if it is among the largest available corpora of its kind, with a total amount of 1.95 million transcribed tokens, it is still a relatively small corpus. Corpus-based methods, which up to now have been used in lexicography on large amounts of written German, need to be looked at in a new way.

However, FOLK is still being set up and will grow further over the project period. The coverage of different interaction domains as well as the coverage of speakers from different regions in Germany and of additional metadata will therefore be constantly improved and expanded over the coming years. Thus, the LeGeDe project works with the most adequate corpus for the analysis of the lexicon of spoken German on an interactional basis. Since lexicographic resources for the German language have not yet been developed for spoken language data, an important task of the LeGeDe project is to develop new approaches to the corpus-assisted analysis of interactional data. A particular challenge is to unite the methods of conversational with those of lexicological and lexicographical analysis.

4. Quantitative corpus analysis

One of the challenges of the LeGeDe project is to develop automatic, semi-automatic and manual analysis methods, which serve different purposes: The results of automatic methods are used to pre-structure data sets related to different areas e.g. information about combinatorics, formal realisation and meta linguistic data, so that they can be used for the lexicographic resource and be commented on by the lexicographers. The editorial elaboration of the dictionary entries is, of course, another important part of the project work, but this paper does not elaborate on this point.

The linguistic units to be included in the lexicographic resource should, above all, satisfy the criterion of having relevance in the spoken language. Wherever it is meaningfully possible, the aspect of distinctiveness should be taken into account in comparison to written German. In order to assist the detection of salient terms in spoken German we work with frequency comparison between FOLK and DEREKO ("Deutsches Referenzkorpus", written German; eng.: German reference corpus[6]). DEREKO (cf. Kupietz/Keibel, 2009) is much larger: it currently comprises about 29 billion running words. Our assumption is that noticeable frequency differences may indicate to differences in meaning and use. We apply different measures for frequency comparisons, such as Log Likelihood Ratio (Dunning, 1993), Odds Ratio and frequency classes (Perkuhn et al., 2012). The comparative analyses with DEREKO, as a corpus with a wide coverage of many different types of texts, are limited to a subset of the data. For instance, we excluded the Wikipedia sources because of the conceptually spoken German used in the discussion pages. Since DEREKO and FOLK differ in corpus sizes (DEREKO = 29 billion text words vs. FOLK = 202 h / 1.95 million tokens) and temporal coverage of the sources (DEREKO = 1772-2015 vs. FOLK = 2003-2016) differences in metadata and text types must be judged very carefully between the two corpora. They should serve as a frequency-controlled aid to interpretation (see for example the article by Kupietz/Schmidt (2015) on the written and oral corpora at IDS as the basis for empirical research).

After the frequency comparison of the two corpora, we identified different lexical units of interest, such as verbs (*gucken*, *kriegen*, *finden*, *meinen* etc.), particles in the broad sense (*mal*, *halt*, *eben*, *ah*, *oh*, *okay* etc.), adjectives (*gut*, *prima*, *schön*, *geil*, *krass* etc.), nouns (*Ding*, *Sache*, *Stress* etc.), and pronouns (*etwas*, *was*, *solch-*, *irgend-* etc.). An excerpt of the table for frequency analysis representing the particles with the highest difference in frequency classes can be observed in Table 3.

Lemma	**FOLK absolute frequency**	**DEREKO absolute frequency**	**FOLK frequency class**	**DEREKO frequency class**	**Difference of frequency class**
okay	6477	199942	4	14	10
halt	6136	802658	4	12	8
mal	14076	8523173	2	8	6
na	3077	520673	5	12	7

Table 3: Frequency comparisons: particles (excerpt)

We also use the comparison of frequency classes for studying the distributional behaviour of pseudo-synonyms, such as between the verbs *gucken* and *schauen* (see Table 4).

[6] Information about DEREKO: http://www.ids-mannheim.de/kl/projekte/korpora/.

Lemma	FOLK absolute frequency	DeReKo absolute frequency	FOLK frequency class	DeReKo frequency class	Difference of frequency class
gucken	2598	375327	5	13	8
schauen	570	2570951	7	10	3

Table 4: Frequency comparisons: *gucken* vs. *schauen* (excerpt)

In addition, since we categorised all the transcripts in FOLK into interaction domains such as "private", "public", "non-private/non-public" and "other" (see section 3, Figure 1), we determine the distribution of lexical items within different categories. Such an indication can refer to a single element (example *gucken*), but it can also be considered in relation to the distribution of all lemmas in FOLK. We also use this categorisation in order to study the lexical units belonging to the same phenomenon class (example: visual perception verbs; *gucken*, *schauen*, *sehen*; cf. Figure 2).

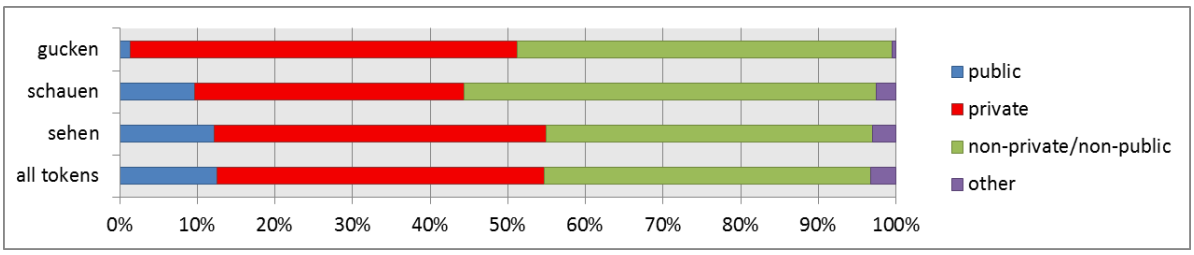

Figure 2: Distribution on different interaction domains. Comparison: visual perception verbs (*gucken*, *schauen*, *sehen)* - total amount of all tokens

The comparison in Figure 2 shows on the one hand, that the frequency of the verb *gucken* is relatively higher in private conversations compared to the other two visual perceptual verbs (*schauen* and *sehen*); in addition, *gucken* is much less common in public conversations. On the other hand, compared to all tokens in FOLK, *gucken* rarely occurs in public conversations and with increased frequency in private contexts.

Since our first case studies focus on verbs, in order to obtain a fine-grained analysis of the verb distribution in FOLK, we perform a reconstruction of separable particle verbs in the corpus (Volk et al., 2016; Batinić/Schmidt 2017). In that way, verbs such as *angucken* or *anschauen* can be extracted from the corpus even when they are not written together, a piece of information usually not available in the default lemmatisation of most corpora. Since FOLK contains not only transcribed words, but also their normalised and lemmatised forms, we can perform frequency measurement on each formalisation level. In order to have an overview of the word form frequencies on each level, we produce a word profile containing the frequency of transcribed word forms for each annotation level (cf. Table 5).

Lemma	Norm	Transcription
gucken	*geguckt*	*geguckt* 81, *gekuckt* 2, *geguck* 2
gucken	*gucken*	*gucken* 686, *gucke* 77, *gugge* 34, *kucken* 28, *guckn* 7, *guck* 5, *gu* 5, *kucke* 4, *kuck* 3
gucken	*guckten*	*guckten* 2
gucken	*guckte*	*guckte* 3
gucken	*gucke*	*guck* 105, *gucke* 28, *kuck* 22
gucken	*guckt*	*guckt* 111, *kuckt* 6, *guck* 3
gucken	*guckst*	*guckst* 79, *gucks* 33, *gucksch* 4, *kuckst* 3, *guckscht* 2
gucken	*guck*	*guck* 475, *gu* 82, *kuck* 13, *ku* 10, *gugg* 8, *gucke* 2, *kiek* 2

Table 5: Frequency of transcribed word forms
for each annotation level (example *gucken*)

We also study word distributions by using different meta-information about region and speaker. Table 6 shows selected words that are less frequently used by men as by women.

Lemma	Male (948586 tokens)	Female (980190 tokens)	Range (number of speakers)	Log Likelihood	Odds Ratio
Gott	212	598	214	179,20	0,37
ups	17	87	48	49,27	0,20
juhu	6	47	19	34,71	0,13
boah	148	380	162	98,04	0,40

Table 6: Distribution via the parameter "gender" (excerpt)

In addition to analyse one word lemmas, we also focus on multiword expressions. We identify frequent words that co-occur with the target word as well as the most frequent bi- and tri-grams containing the target word (we work with absolute frequencies given the relatively small size of the corpus). The co-occurrence profiles are commonly used for the analysis of corpora of written language (for the creation and use of word profiles in lexicography see e.g. Adam Kilgarriff's work on Word Sketches: e.g. Kilgarriff & Kosem, 2012 or Kilgarriff, 2015). These methods have not yet been applied to data material for spoken German, especially with regard to FOLK. Missing sentence boundaries, speaker changes, uncertain word forms, and overlaps, etc. are only a few challenges in this regard. The project deals with the opportunities and limitations of such statistical procedures.

After detecting salient word combinations (e.g. *guck mal*, *müssen wir mal gucken*) we analyse them in detail in the coding part (see section 5). An overview of some frequent co-occurrences (word combinations, patterns, etc.) of the verb *gucken* is shown in Figure 3.

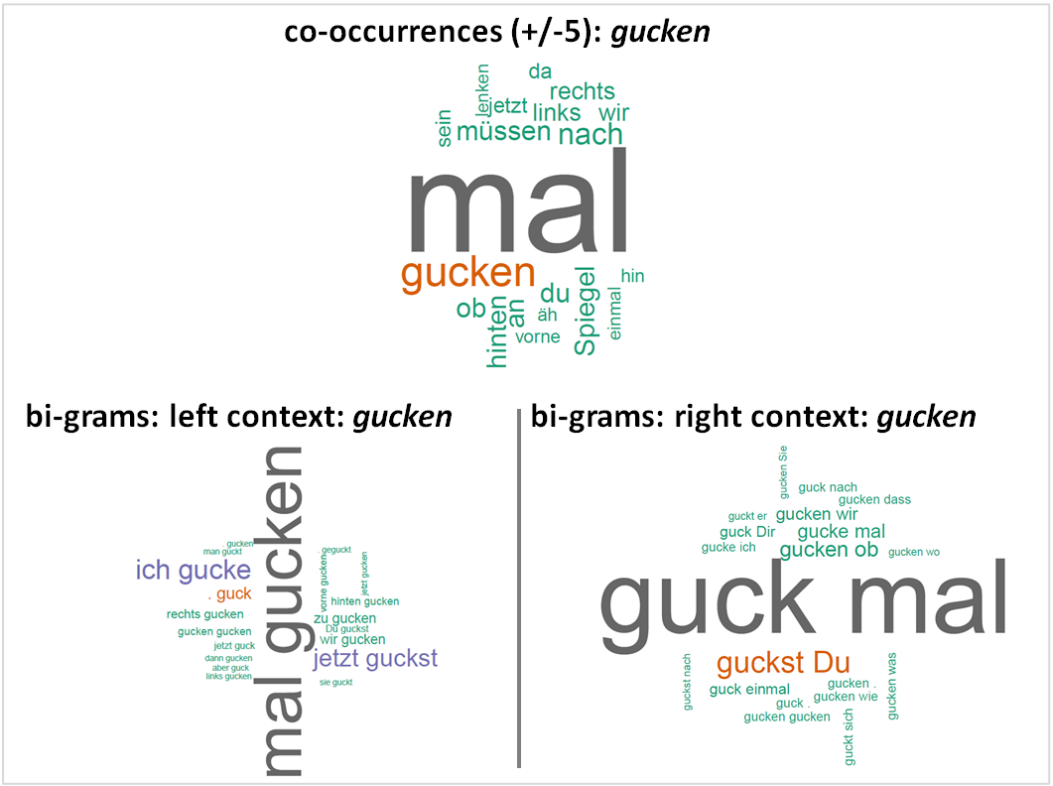

Figure 3: Co-occurrences and bi-grams with regard to the verb *gucken*

5. Data analysis

We have carried out the first in-depth analyses with verbs, which we exemplarily illustrate in this section. The first steps (sampling, creation of a coding table) involved the elaboration of a coding scheme as well as the analysis and structuring of the data – especially in connection with initial considerations about the development of a lexicographic microstructure.

In order to extract corpus samples constraining a particular lemma, we defined following preliminary steps: a) assigning all conversations to 4 different interaction domains ("private", "public", "non-public/non-private", "other"; see Table 2 and Figure 1), b) calculating the distribution of the lemma to the interaction domains with regard to the whole corpus and c) transferring the distribution to the proportion with regard to the sample.

METADATEN TREFFER				
SHORT_FILE ID	LINKER KONTEXT	STICHWORT	RECHTER KONTEXT	LINKID
FOLK_E_0024	weiß ich nich wie ich	**gucke**		http://dg
FOLK_E_0002	dann	**guck**	isch heut ma was wir morgen dann noch einkaufen gehen müssen	http://dg
FOLK_E_0002	ja mir müsse mal	**gucke**	was mer für wichtige halte und was net vielleicht kannscht ja afach da irgendwie n hake dra mache	http://dg
FOLK_E_0008	ja	**guck**	mal die arbeitsumstände die werden ja immer dem land angepasst in dem du dann grade deine filiale aufmachst	http://dg
FOLK_E_0020	is immer so ich geh dann einkaufen un dann will ich ja eintlich gar nisch aber dann wenn ich immer so drinnen bin dann kann ich irgendwie dann immer so schön	**gucken**	un da guck ich immer mit	http://dg

Figure 4: Extract from an excel spreadsheet of the search results to *gucken* (eng. *to look*) (FOLK, DGD)

Each KWIC line of our sample has a column with a link to the corresponding transcript excerpt in the database (see Figure 4; DGD, FOLK). In this way, the larger context of an occurrence and the corresponding audio recording can be inspected, both of which are essential for the various analysis steps (see Figure 5).

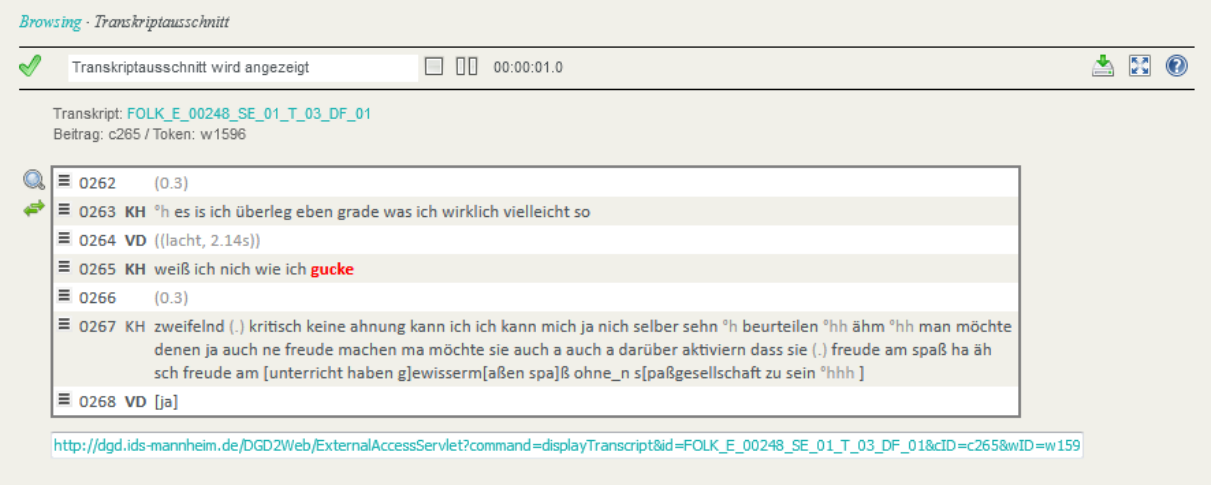

Figure 5: Corpus reference to the link from the excel sheet to the verb *gucken*, KWIC line 1

For coding the data, a coding scheme has been developed for 5 different coding areas with different coding parameters (see Figure 6). In addition to the different automatically generated metadata regarding the hit itself (section 1), there is an automatically generated information on meta-language data concerning the transcript (section 5). The data are examined through "hands-on analysis", with regard to content-functional analysis (section 2), syntactic-formal analysis (section 3) and

grammatical information (section 4).

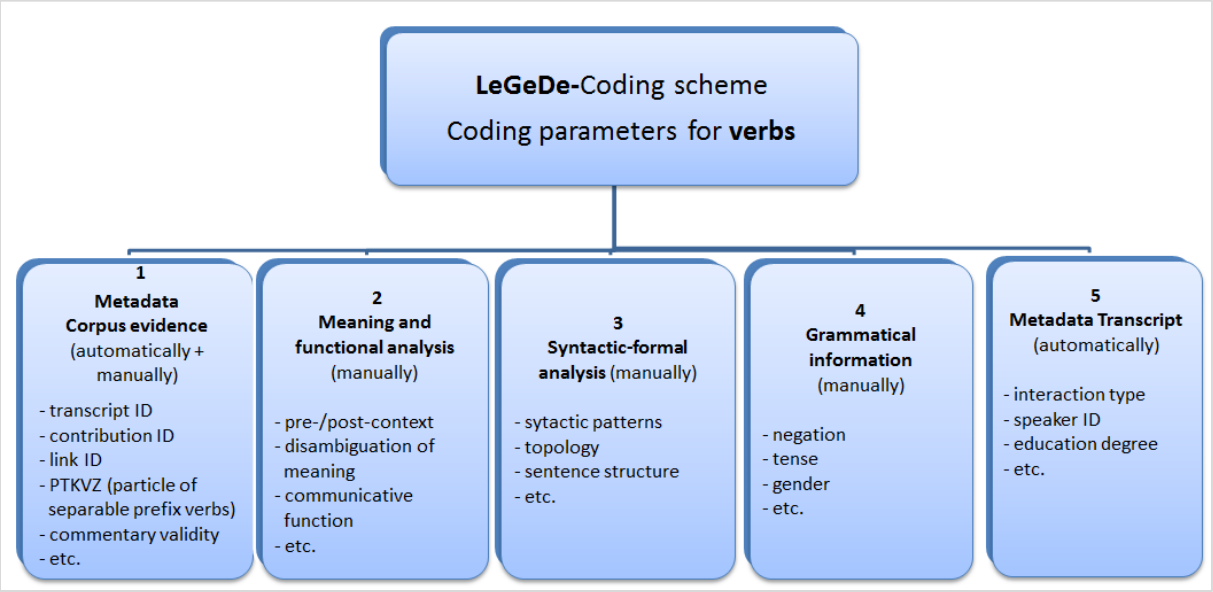

Figure 6: Coding parameters for verbs

The coding scheme is continuously refined in several encoding processes, which are carried out by several persons. Multiple encoding processes and different persons' examinations of the data are intended to increase precision in the coding and interpretation of the data, particularly in the meaning-disambiguation and the differentiation of the function of a word or a phrase in the interactional context.

As already mentioned in section 1, the description of the peculiarities, especially in the area of the lexis of spoken German, is only inadequately documented in existing dictionaries. Figure 7 shows an extract of the dictionary article *gucken* from one of the most consulted Learner dictionary for German as a foreign Language (LGWB-DaF). The extract from the dictionary article shows grammatical information (verb intransitive, sentence structure patterns, ["irgendwohin / irgendwie gucken…"]) and information on the meaning (definition, paradigmatic relations). The dictionary user also finds the very general pragmatic information that the lemma *gucken* is a lemma used in contexts of spoken German (label: "gesprochen"). Only three meanings of the lemma *gucken* are listed in this dictionary.[7]

[7] In the Pons Kompaktwörterbuch (Deutsch als Fremdsprache – German as a foreign language; 2016), two meanings are listed, the Duden 10 (Bedeutungswörterbuch - explanatory dictionary; 4th edition 2010) and the website of Duden-online show three different meanings of *gucken*.

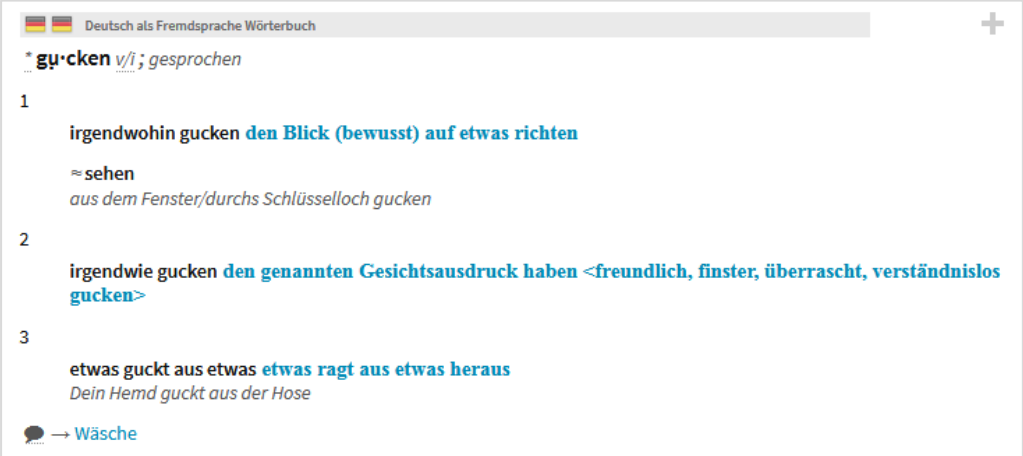

Figure 7: Extract from the dictionary article *gucken* from the "LGWB-DaF"

Our analyses of the lemma *gucken* indicate that we have come to a more expanded understanding of the meanings, formal realizations, and ultimately of the function of the verb *gucken* compared to information from standard German dictionaries and, particularly, of learners' dictionaries. According to our investigations, the spectrum regarding the meanings of *gucken* is much larger. We performed the semantic disambiguation by analyzing the form (" [argument] structure pattern" in conjunction with the corresponding "sentence structure") and content (cf. Table 7).

Semantic definition / meaning	**Synonyms**	**(= STM) (argument) structure pattern**	**(= SBP) sentence structure**[8]
…	…	…	…
jmd. stellt fest, dass etw. d. Fall ist	feststellen	jemand *guckt*, dass etwas der Fall ist	<Ksub, Kverb>
jmd. sieht s. etw. an	sich ansehen	jemand *guckt* etwas	<Ksub, Kakk>
jmd. beobachtet, wie etwas passiert	beobachten zuschauen	jemand *guckt*, wie etwas passiert	<Ksub, Kverb>
jmd. sucht nach etwas	suchen	jemand *guckt* nach etwas	<Ksub, $Kprp_{nach}$>
jmd. schaut sich um	umherschauen	jemand *guckt* auf eine bestimmte Art und Weise	<Ksub, Kmod>
jmd. passt auf, dass etwas (nicht) passiert	aufpassen kontrollieren	jemand *guckt* dass etwas (nicht) passiert	<Ksub, Kverb>
…	…	…	…

Table 7: Different meanings of the lemma *gucken* (excerpt)

[8] Terminology in accordance with Zifonun et al., 1997.

As FOLK forms our data base, it is possible for us to work especially on interaction-specific information and to implement it for the planned lexicographic resource. The following information would be interesting and could complement the offer of existing dictionaries profitably: the interaction context or sequence context, prosody and sound realisation, large variety in functional aspects with regards to the interaction context, combination potential (cf. Figure 3 in section 4 and the discussion about automatically generated co-occurrence profiles and the identification of combination potential), information about topology, and other aspects.

Taking into account the corresponding interaction context and the metadata, conclusions can be drawn about the respective possibilities of use and the corresponding communicative functions. With FOLK as a data base, the expertise in the project on conversational analysis as well as the expertise in the field of lexicology and lexicography, the project would like to close the gap with respect to the interaction-specific information for verbs as well as for other word classes and lexical patterns.

6. Final remarks

During the project period we want to develop corpus-based methods for analyzing and structuring spoken lexis as well as a lexicographical process that take the characteristics of language in interaction and the possibilities of the database into account. The sub-targets of the project can be described as follows: (i) determination of the peculiarities and divergences of the spoken and written language usage in the lexical area at all levels (form, content / function, situation etc.), (ii) development of further corpus linguistic methods for analysing and structuring the data of spoken language, (iii) development of innovative types of lexicographical information, which refer to the function of lexical units in interaction contexts, (iv) development of innovative description formats in a multimedia format for lexical data. The aim is to offer the user a mixture of automatically generated data (see section 5 in particular) as well as lexicographically commented information (see section 6 with regard to the analysis steps).

The lexicographically commented information will include aspects such as peculiarities in form (form-related realization, word forms, inflection, phonetic realization, etc.), combinatorics (actants, morphosyntactic information, etc.), meaning (meaning description, conceptual reference, paradigmatic sense relations, etc.) and communicative function (combination of topology, formal aspects, interactional criteria, metadata, etc.). From the lexicons' specifics in oral communication, new challenges arise for the macro, micro and medio structure of this new type of dictionary, as well as for an electronic presentation that must combine text with multimedia forms of expressions.

Besides being used for linguistic research, the lexical resource could contribute to the acquisition of German as a foreign or second language and as well as to the development of a language-reflexive first language teaching[9].

The LeGeDe project not only contributes to a new description of contemporary German, but also to the development of lexical descriptions appropriate for the lexis of spoken German. The lexicographic resource is intended to describe the lexical competences of everyday conversation and to contribute to the better understanding of the peculiarities of the vocabulary of spoken German in interaction.

7. Acknowledgements

We would especially like to thank our colleagues Rainer Perkuhn, Cyril Belica and Marc Kupietz (program area "Corpus Linguistics" at the IDS), which support us in many corpus linguistic questions. They are of great help to the LeGeDe project in aspects relating to the corpora of spoken German and the analysis procedure for these corpora. We also thank our colleagues from the Department of Pragmatics and the Department of Lexical Studies, who have actively supported us with their advice and ideas. Our thanks go to Henrike Helmer, Julia Kaiser, Frank Michaelis, Carolin Müller-Spitzer, Nadine Proske and Arne Zeschel.

[9] See e.g. keywords in "Kultusministerkonferenz" [2012: 12]: "Sprache und Sprachgebrauch reflektieren/ Reflecting language and language usage" as well as "Sich mit Texten und Medien auseinandersetzen/Dealing with texts and media".

8. References

Batinić, D. & Schmidt, T. (2017). Reconstructing of Separable Particle Verbs in a Corpus of Spoken German. To appear in *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2017)*, September 13 – 15, 2017, Humboldt Universität zu Berlin, Germany.

Deppermann, A. (2005). Conversational interpretation of lexical items and conversational contrasting. In A. Hakulinen & M. Selting (eds.) *Syntax and lexis in conversation*. Amsterdam: Benjamins, pp. 289–317.

Deppermann, A. (2007). *Grammatik und Semantik aus gesprächsanalytischer Sicht*. Berlin: de Gruyter.

Deppermann, A. & Proske, N. & Zeschel, A. (eds.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincide. In *Computational Linguistics*. 19 (1), pp. 61–74.

Fiehler, R. (2016). Gesprochene Sprache. In A. Wöllstein & Dudenredaktion (eds.) *Duden – Die Grammatik.* Berlin: Dudenverlag, pp. 1181–1260.

Günthner, S. (2016). Diskursmarker in der Interaktion – Formen und Funktionen univerbierter *guck mal*- und *weißt du*-Konstruktionen. In *SpIn-Arbeitspapierreihe (Sprache und Interaktion)*, Nr. 68.

Hansen, C. & Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012*. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929–935.

Imo, W. (2007). *Construction Grammar und Gesprochene-Sprache-Forschung. Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch*. Tübingen: Niemeyer (= Germanistische Linguistik, Band 275).

Kilgarriff, A. & Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: Oxford Univ. Press, pp. 31–55.

Kilgarriff, A. (2015). Using Corpora as Data Sources for Dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography.* London/New Delhi/New York/Sydney: Bloomsbury, pp. 77–96.

Kulturministerkonferenz (2012). *Bildungsstandards im Fach Deutsch für die allgemeine Hochschulreife*. (Beschluss der Kultusministerkonferenz vom 18.10.2012).
http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf. (Accessed at: 10 July 2017).

Kupietz, M. & Keibel, H. (2009). The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In *Working Papers in Corpus-based Linguistics and Language Education, No. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53–59.

Kupietz, M. & Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L. Eichinger (ed.) *Sprachwissenschaft im Fokus*. Berlin/Boston: de Gruyter, pp. 297–322. (= Jahrbuch des Instituts für Deutsche Sprache 2015).

LGWB-DaF: *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (Neubearbeitung 2015; online access via the IDS library).

Meliss, M. (2016). Gesprochene Sprache in DaF-Lernerwörterbüchern. In B. Handwerker & R. Bäuerle, & B. Sieberg (eds.) *Gesprochene Fremdsprache Deutsch*. Baltmannsweiler: Schneider, pp. 179–199. (= Perspektiven Deutsch als Fremdsprache, Band 32).

Perkuhn, R. & Keibel, H. & Kupietz, M. (2012). *Korpuslinguistik*. (= UTB 3433). Paderborn: Fink.

Schmidt, T. (2014a). The Research and Teaching Corpus of Spoken German - FOLK. In *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA, pp. 383–387.

Schmidt, T. (2014b). The Database for Spoken German - DGD2. In *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA, pp. 1451–1457.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. In J. M. Kirk & G. Andersen (eds.) *Compilation, transcription, markup and annotation of spoken corpora, Special Issue of the International Journal of Corpus Linguistics* [IJCL 21:3], pp. 396–418.

Schwitalla, J. (2012). *Gesprochenes Deutsch. Eine Einführung*. 4., neu bearbeitete und erweiterte Auflage. Berlin: Schmidt. (Grundlagen der Germanistik 33).

Verdonik, D. & Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In *International Journal of Lexicography.* 30 (2), pp. 143–166.

Volk, M. & Clematide, S. & Graën, J. & Ströbel, P. (2016). Bi-particle adverbs, PoS-tagging and the recognition of german separable prefix verbs. In: KONVENS 2016, Bochum, 19-21 September 2016. https://doi.org/10.5167/uzh-126372. (Accessed at: 10 July 2017).

Westpfahl, S. & Schmidt, T. (2016). FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16), Portorož, Slovenia*. Paris: European Language Resources Association (ELRA), pp. 1493–1499.

Zifonun, G. & Hoffmann, L. & Strecker, B. (1997). *Grammatik der deutschen Sprache*. Berlin/New York: de Gruyter. 3 volumes.

Dictionaries and dictionary portals:

Duden 10 - Das Bedeutungswörterbuch (2010). Mannheim: Duden-Verlag.

Duden-online. Accessed at: http://www.duden.de. (10 July 2017).

DWDS: *Digitales Wörterbuch der Deutschen Sprache*. Accessed at: https://www.dwds.de. (10 July 2017).

elexiko. Accessed at: http://www.elexiko.de. (10 July 2017).

LGWB-DaF: Langenscheidt Großwörterbuch Deutsch als Fremdsprache (2015).

München: Langenscheidt.
OWID. Accessed at: www.owid.de. (10 July 2017).
Pons Kompaktwörterbuch Deutsch als Fremdsprache (2016). Stuttgart: Pons.

Websites:

Datenbank für Gesprochenes Deutsch. Accessed at: http://dgd.ids-mannheim.de. (10 July 2017).
DeReKo. Accessed at: http://www.ids-mannheim.de/kl/projekte/korpora/. (10 July 2017).
FOLK (Information about the corpus). Accessed at: http://agd.ids-mannheim.de/folk.shtml. (10 July 2017).
Leibniz-Gemeinschaft. http://www.leibniz-gemeinschaft.de/en/about-us/leibniz-competition/projekte-2016/funding-line-1/. (10 July 2017).
Lexik des gesprochenen Deutsch. Information about the project: http://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html. (10 July 2017).